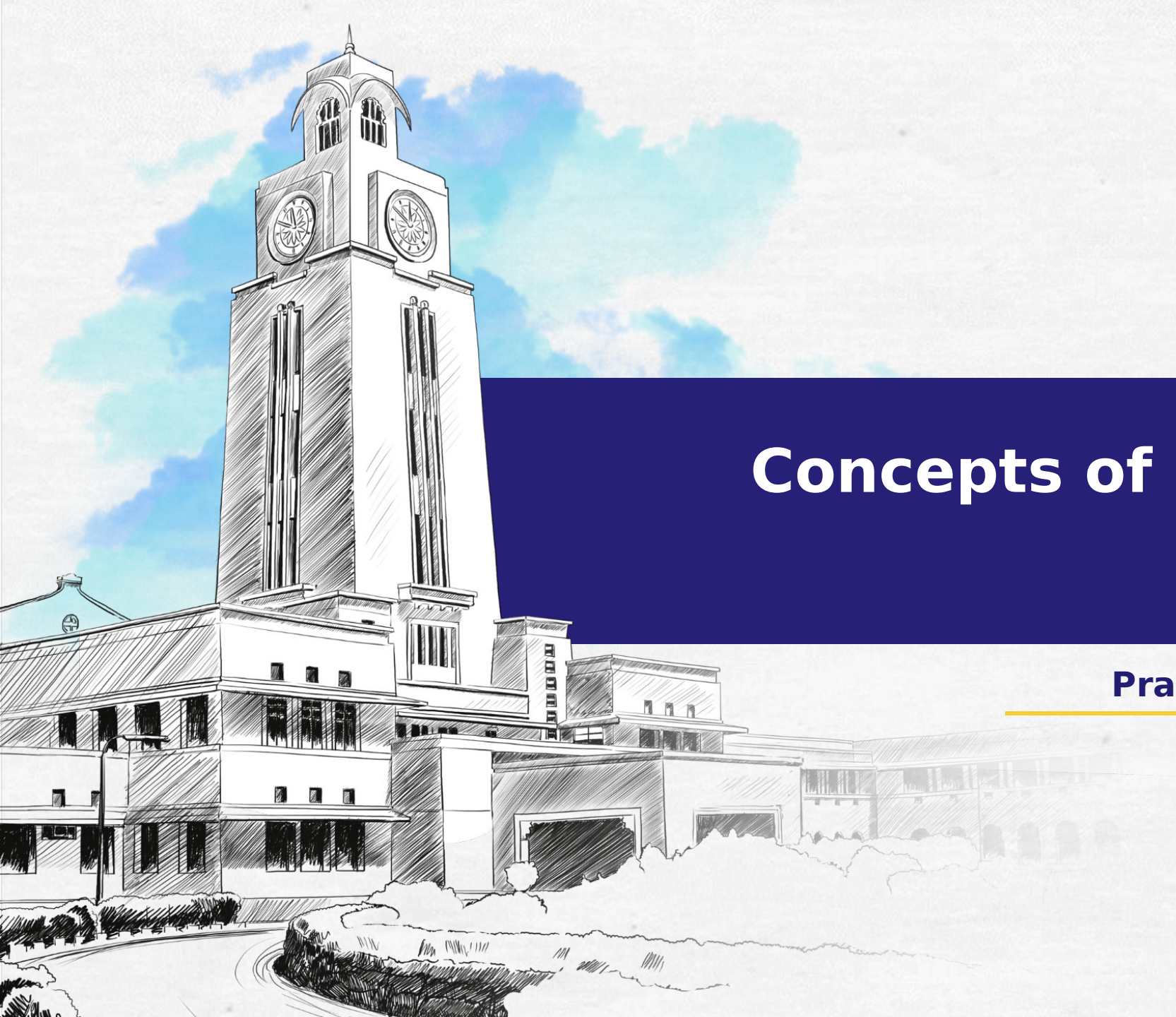




BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Concepts of Data Management

Pravin Y Pawar & Tharabai Baboo

Agenda

- Principles of Data
- Data Management
- Data Platform
- DM Components
- Phase of data life cycle
- DM Roles
- Connecting Dots





Data As Asset

Data as
Asset

Data as
Liability

Data Sensitivity of ML
Pipelines

Data Reliability



Trade-offs of data size, model error rates, and risk of problems or issues associated with the data

What this graph says? Your views?

Can you imagine a ML system without data?

Case Study: Netflix
(<https://research.netflix.com/research-area/analytics>)

“we are able to provide efficient and actionable data-informed perspectives that help Netflix think critically and differently about its business and ultimately improve our service”

What this implies?

- Data is an important asset
- For a successful business model, data is key
- Transform data to value for successful

Data As Liability

Data as
Asset

Data as
Liability

Data Sensitivity of ML
Pipelines

Data Reliability

Very Basics of Data Management



Problem Statement:

- One man's food is other man's poison
 - Similarly anything can be an asset or **liability** basis the circumstances
 - **Holds good for data as well**

How Data can become liability?

- Exposure to unexpected nuances of collected data
- Not considering this can harm us as well as end users

Example:

- Medical records – can be an asset or liability if exposed unnecessarily

“In the context of healthcare, the potential legal and ethical risks associated with storing and managing sensitive patient medical records, where a breach or misuse of this data could lead to significant consequences for the healthcare provider, including lawsuits and reputational damage; essentially, the sensitive nature of medical information means that mishandling it can create legal liability for the entity responsible for its storage and access.”

Data As Liability

The biggest Myth

More data ==
better

This stuff is easy

Changes at

Phase

You would have seen this, when applying for loans/credit cards, etc.

☐ I hereby accept that my Personal Data may be held and processed by the bank for availing this loan and relevant services. I Agree to Perfios T&C

☐ I hereby allow the bank to process my data for sending me personalized offers on relevant products and services

☐ I allow the bank to reach out to me with personalized offers on new products through SMS, WhatsApp, Email and Voice

TechnoFino

Data as
Asset

Data as
Liability

Data Sensitivity of ML
Pipelines

Data Reliability

Collect

Store

Delete

Restrictions In Data Collection

Compliances:

- Location of organization (Eg: In employer can store employees' data in In servers)
- Location of data origin(Eg: European Union data cannot exit the region)
- Organization policies (Eg: Confidential data of the organization)

Restrictions:

- Data about people(PII?)
- Permission to store data(how?)
- Storage(how?)
- Retrieval of permission(how?)
- Access to the data provider(how?, when? and why?)

Sources of restrictions:

- Governing laws
- Industry practices
- Insurance regulations

A Classic example of collecting PII(Personally Identifiable Information) - Per law, this needs an explicit consent along with

Data As Liability

The biggest Myth

More data ==
better

This stuff is easy

Challenges at Phase

Approach 1 : Access Restriction

- Access restrictions(external and internal)
- Restrict even employees accessing or modifying the private user data. If needed then,
 - ✓ Impose granular restrictions
 - ✓ Enable detailed logging to trace back

Approach 2 : Anonymization of Data

- To reduce the data access and auditing surface
- Pseudonymization
Eg: Postcode. For weather predictions, a diff post code from the same locality can be given.
- Replace private identifiers with others in a reversible fashion
- Reversing needs access to additional data or systems
- Protects data from casual inspection by engineers along the pipeline
- Preserves the properties of data that are relevant to the model(Eg: Postcode from same locality)

Approach 3 (Recommended) : Connection Removal

- Remove connection between private data(PII) and model data
- Reduces the risk
- Harder than it seems
- Not trivially reversible but still valuable can be difficult

Data as
Asset

Data as
Liability

Data Sensitivity of ML
Pipelines

Data Reliability

Collect

Store

Delete

Data As Liability

The biggest Myth

More data ==
better

This stuff is easy

Challenges at Phase

Data as
Asset

Data as
Liability

Data Sensitivity of ML
Pipelines

Data Reliability

Collect

Store

Delete

➤ When will you delete data?

- ✓ When the user mentions a specific time period
- ✓ When the org no longer need that data

Eg:

Have you noticed companies' job portals which says that your data/profile details will be retained for 6 months only?

After that you have to recreate your profile in the companies' job portal.

- ✓ When the governing policies sets a cut off(Eg: GDPR)

➤ Harder than we speak

➤ When you say Shift+Del - what happens the actual data ?

➤ How the recovery softwares can restore the data in any device?

➤ Eg:

Consider a system where data is written on to tape drive for archival purpose. Now deleting data in primary system is possible, but how will you delete data from tapes?

Data Sensitivity of ML Pipelines

Data as
Asset

Data as
Liability

Data Sensitivity of ML
Pipelines

Data Reliability

Scenario:

An ecommerce company wants to predict the new year eve sales, ie sales on Dec 31st, from all the countries, region or language.

They also want to study some trend based on the language.

Now due to some technical or human error, they lost data from a particular country or region or language.

The data pipeline was still executing normally while ML pipeline was behaving so differently.

Can you reason out why?

- Data pipeline is not predicting anything, so it can work as usual
- But ML pipeline includes the model as well
- So if the right data is not supplied then ML pipelines behave abnormally bringing down the performance as well as the accuracy
- Now in the scenario, any of the omission can result in incorrect prediction and impact the sales and reputation of the ecommerce company
- Normal data pipelines are sensitive to
 - ✓ Volume of input data
 - ✓ Correctness of input data
- ML Pipelines are sensitive to
 - ✓ Changes to distribution of data

Data Reliability

Data as
Asset

Data as
Liability

Data Sensitivity of ML
Pipelines

Data Reliability

In the growing distributed environment, what are the minimum characteristics expected from data?

After some time, what will happen if the needed data is lost?

Durability

Are we not supposed to get same data from all its copies?

Consistency

In a team, where the data changes frequently, don't we want to time travel?

Version control

How fast the data should be available to get accurate answers?

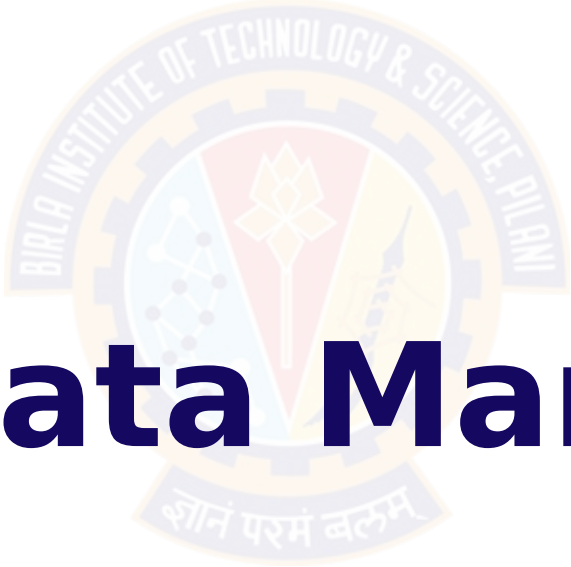
Performance

As and when we need is the data available?

Availability

Leads to the character

Home Work: Find out why integrity is not there

The logo of Birla Institute of Technology & Science (BITS) is a circular emblem. The outer ring contains the text "BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI" in a serif font. The inner circle is divided into three colored segments: a blue segment on the left with a network diagram, a red segment in the center with a stylized lotus flower, and a yellow segment on the right with a gear-like pattern. Below the circle is a banner with the Sanskrit motto "ज्ञानं परमं बलम्" (Gyanam Paramam Balam).

Data Management

A study by IBM

What sets leading CEOs apart?

The AI advantage

CEOs are on the hunt for AI investments that will deliver a wide range of business outcomes. The top-performing CEOs in our study are ahead of the curve. They expect AI to have a more transformational impact on business objectives over the next two years.



- 72% of top-performing CEOs agree that competitive advantage depends on who has the [most advanced generative AI](#).

- What this conveys to us?
- In order to take advantage of artificial intelligence (AI), organizations must first organize their information architecture to make their data

Challenges for effective DM?

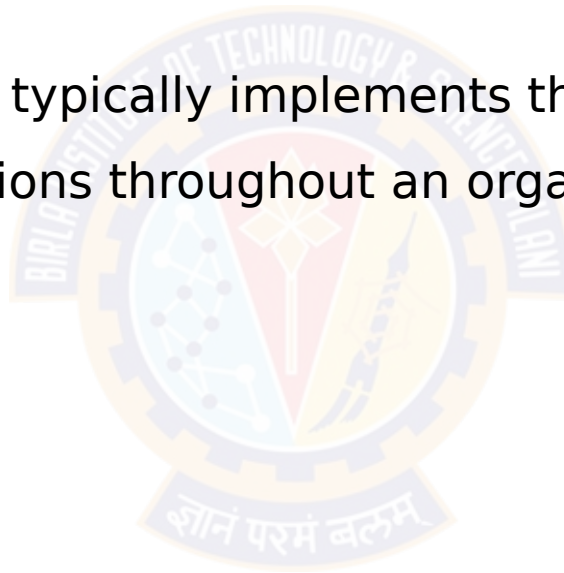
- Data Volume
- Data silos across multiple locations & cloud environments
- Data types and data formats (documents, images, videos)

What is Data Management?

As we already discussed, Data management is the practice of collecting, processing and using data securely and efficiently for better business outcomes.

A strong data management strategy typically implements the following components to streamline their strategy and operations throughout an organization:

- I. Data integration and processing
- II. Data Storage
- III. Data governance
- IV. Data security



I . Data integration and processing

- Step 1 of data management life-cycle
- Raw data ingestion from sources like
 - Web APIs
 - Mobile Apps
 - IOT devices
 - Forms
 - Surveys
- Once ingested, what next to be done?
- Can we use the data as such?

What is this?



The ETL Process Explained



- historically been the standard method to integrate and organize data across different datasets

I. Data integration and processing - contd

- Now what is the problem with ETL?
- With growing trends in data

Are our data always structured? YES / NO

Can ETL cater to the growing velocity? YES / NO

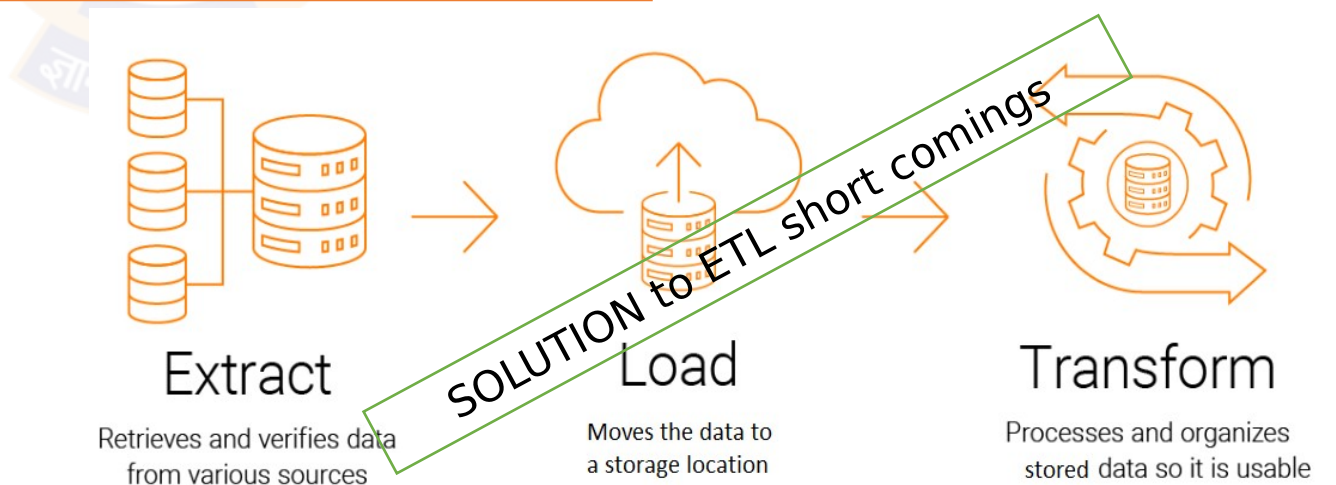
Can it work for real-time data flow? YES / NO

NO

NO

NO

- Popular because of
 - Advancement in cloud infrastructure
 - Need for real-time data



I. Data integration and processing - contd

Can we directly consume the extracted data? YES/NO - still the question is not answered

NO

Next Steps to be done

- Filtered
 - Merged
 - Aggregated
- 
- For what ?
- For the applications that can range from a business intelligence dashboard to a predictive machine learning algorithm

II. Data Storage

When?

Data should / can be stored before / after data processing

Deciders

Type & Purpose of the data



Warehouse



Lake

What do you infer?
Which is structured and which is unstructured here?
Where will you use these?

II. Data Storage

Data Warehouse

- Required a defined schema
- Mostly Relational system or RDBMS
- Optimized for high performance analytics on structured data
- Use cases:
 - Data analytics dashboards
 - Business intelligence tasks
 - Data visualization

Data Lake

- No defined schema.
- Comes handy for semi-structured & unstructured data
- Optimized for costs and multi - type sources
- Use cases:
 - AI
 - Analytics at a scale
 - Sandbox for innovative data projects

+

=

Data Lakehouse

III. Data Governance

Why?

- Promotes :
 - Availability
 - Usage
- Ensures Compliance

- Who does all these?
- Data Governance Councils

What?

- A set of standards and business processes which ensure that data assets are leveraged effectively within an organization

Sample activities of data governance councils

Includes

- Processes
- Policies
- Tools

- They align taxonomies to help ensure that [metadata](#) is added consistently across various data sources

Ensures

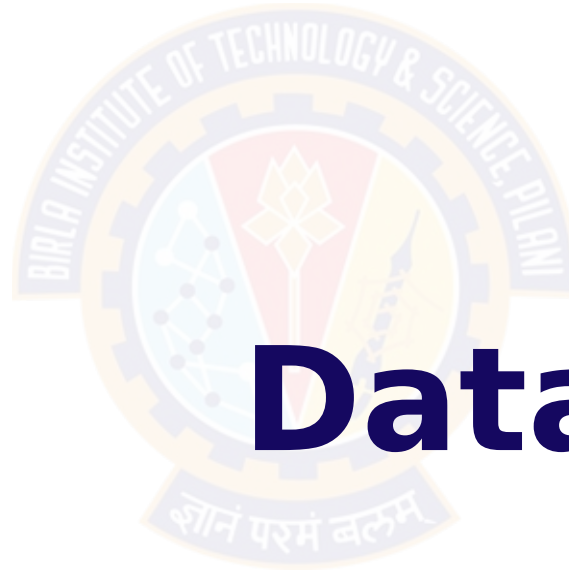
- Data Quality
- Data Access
- Data Usability
- Data Security

- What is metadata?
- Data about data.
- define roles and responsibilities to help ensure that data access is provided appropriately

IV. Data Security

- define roles and responsibilities to help ensure that data access is provided appropriately
- What does this means?
- Can you see the salary slip of your colleague?
- Can you see the medical reports of a fellow patient?
- This means **controlled access is particularly important to maintain data privacy.**
- [Data security](#) sets guardrails in place to protect digital information from unauthorized

- Why?
 - to help protect customer data from cybercriminals or to help prevent incidents that need disaster recovery
 - data loss can be devastating to any business, data breaches, in particular, can result in costly consequences from both a financial and brand standpoint
- How?
 - more scrutiny is placed upon the security practices of modern businesses
 - secure their data by using [encryption](#) and [data masking](#) within their data security strategy



Data Platform

Data Platform

Definitions:

- A software suite that enables organizations to process and manage large amounts of data quickly and efficiently. It manages data collection, normalization, transformation, and application for a given data product.
- A central repository and processing house for all the organization's data.
- A collection of data pipelines that ingest data, store it, transform it, and deliver it to a variety of destinations so the data can be consumed by different enterprise groups for their purposes.

Why:

Data management tools require a robust data and computing infrastructure that scales to meet business demand without impacting performance

To achieve:

Data environments require

fast query processing

large storage capacity

elastic computing

In-memory caches

massively parallel query processing

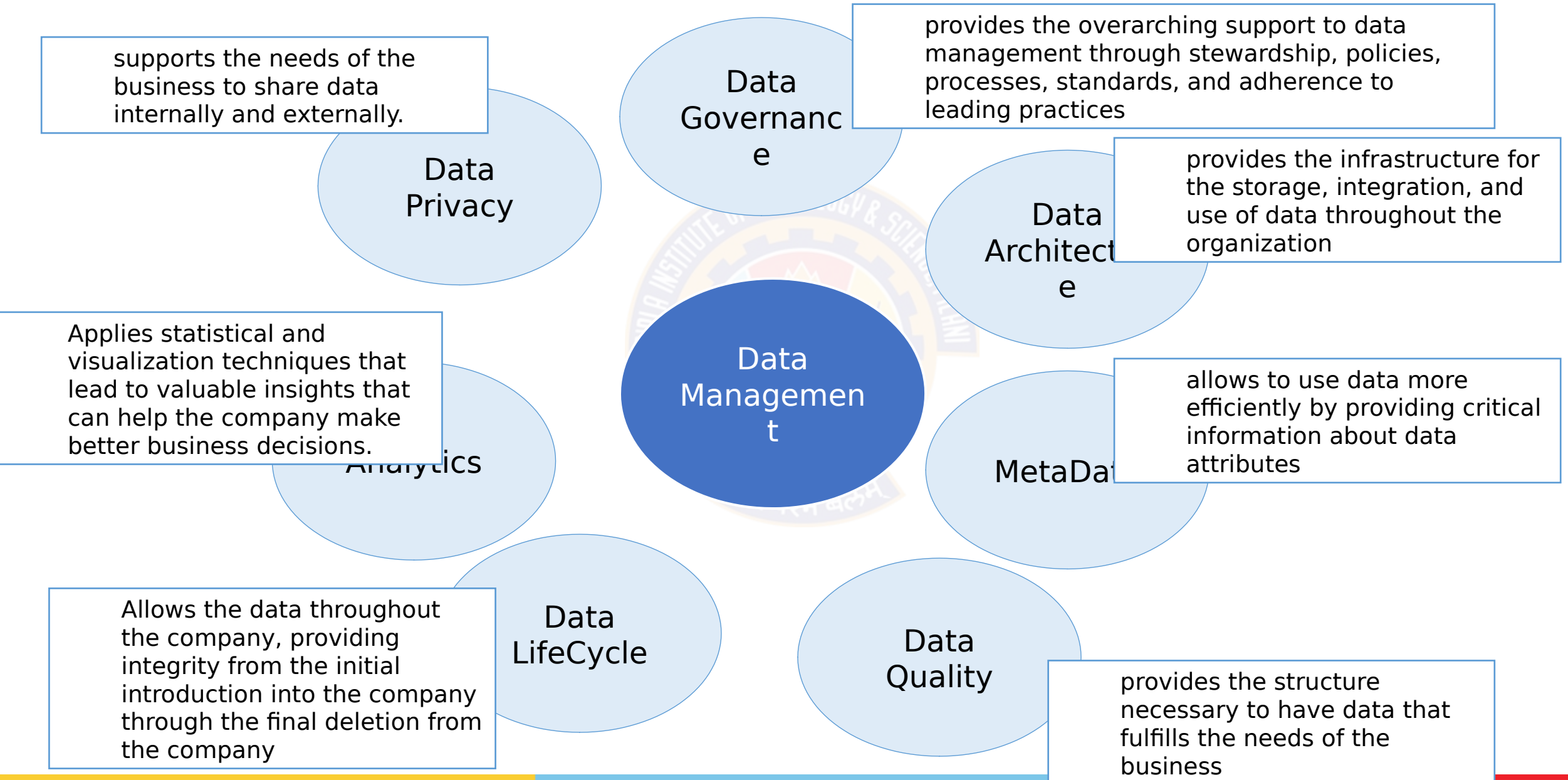
columnar storage

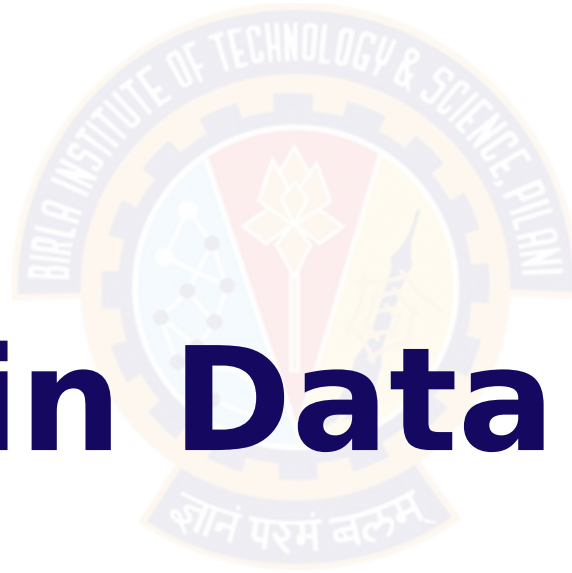
and large in-memory data grids and computing clusters



DM Components

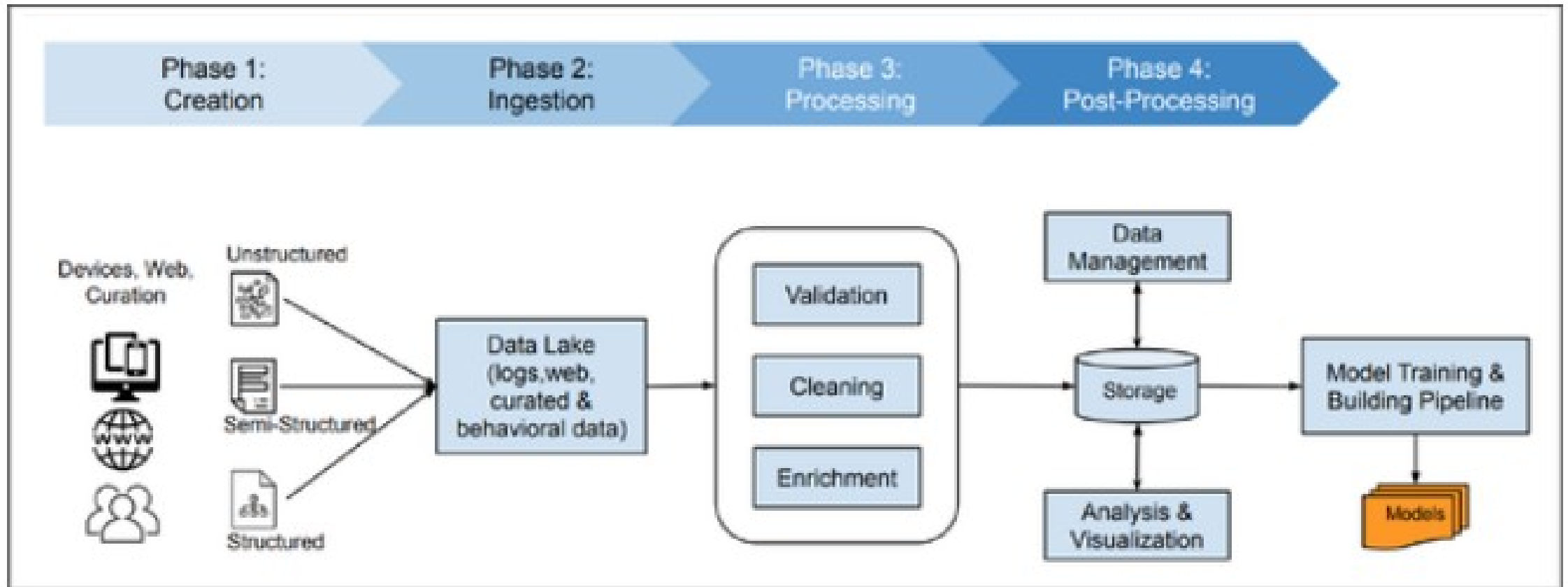
Framework





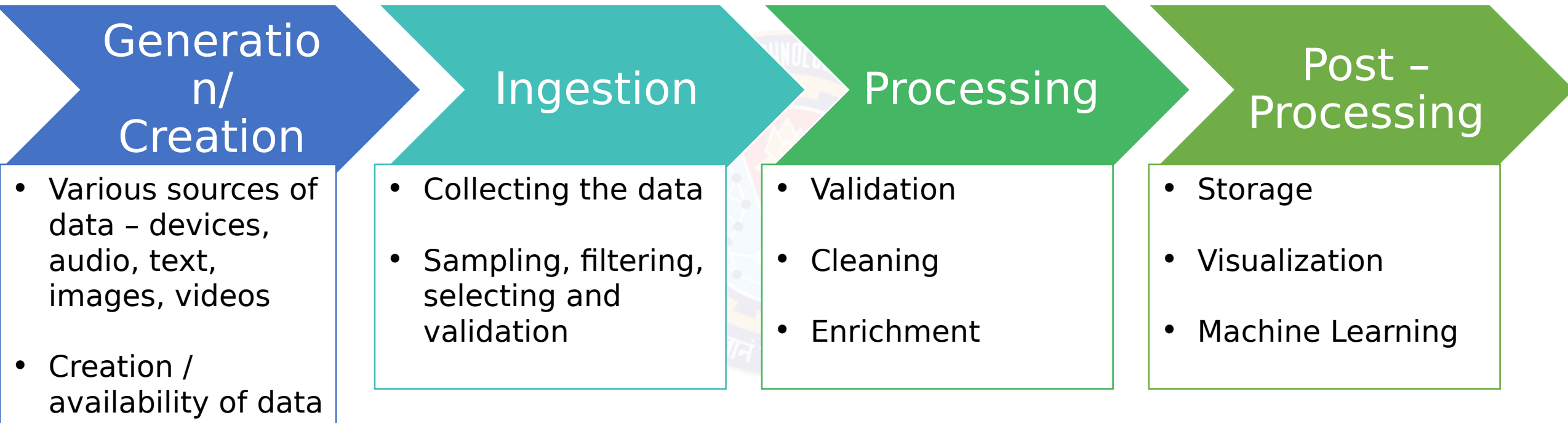
Phases in Data Life-cycle

Phases



ML data management phases

Phases



Generation / Creation

- From where the data comes?

Datasets are all created at some point via some process

- What is data generation or creation?

The process of generating or capturing the data in some data storage system but not our data storage system

- Examples

logs from serving systems, large collections of images captured from an event, diagnostic data from medical procedures

- Some datasets work well when they are static (or at least don't change quickly), & others are useful when frequently updated
 - A photo-recognition dataset, may be usable for many months, as long as it is suitably representative of the types of photos would like to recognize with model.
 - On the other hand, if the outside photos represent only winter environments from a temperate

Ingestion

- The data needs to be received into the system and written to storage for further processing
- Can we take all data that was created / generated at source?
- a filtering and selection step necessarily occurs

Sampling

- may also simply sample at this stage if have so much data which can not be processed
 - an effective way to save money on intermediate processing and training costs
 - important to measure the quality cost of sampling and compare that to the savings
 - is likely to occasionally lose some detail for some events – unavoidable!
- In general, ML training systems perform better with more data. Any reduction in data may well impact quality at the same time as reducing costs.

How?

- Depending on the volume of data and the complexity of our service in the ingestion phase
- APIs / Remote Procedure Calls that gives data in required format and also confirm the data lineage
- Reliability concerns during the ingestion phase typically focus on correctness and throughput

Processing

What will happen to ML if we have bad data?

Solution : Validate the incoming data

- Data ingested from external sources might have a lot of errors
- extremely important to validate the incoming data especially when there is a well-defined schema and/or an ability to compare against the last known valid feed
- performed against a common definition of the field

What will happen if the data has missing fields, duplicates, misclassifications, or even encoding errors?

Solution: Clean them

What if the data is not consistent?

Solution: Normalize the data or bucket the data(map a range of data into a much smaller set of groups that represent the same range)

What if you think the data is not rich enough but available?

Solution : Enrich & extend

- Combine with data from other sources – Enrich and Join
- Identify given event or record – through Labels

Post - processing

Storage:

Should store the processed data somewhere.

How to decide that?

Efficiency of the storage systems

driven by access patterns that are driven by the model structure, team structure, and training process

Metadata - helps humans interact with the storage

- When multiple people work on building models on the same data (or when the same person works on this
- over time), metadata about the stored features provides huge value
- How much importance do you give to detail on metadata?

Machine Learning:

Visualization:

- Process of transforming large amounts of data into an easy-to-navigate representation using statistical and/or graphical techniques and tools
- Not enough just to show a pie chart or bar graph
- Need to provide the human reader an explanation of
 - what each record in the dataset means,
 - how it is linked with the records in other datasets
 - and whether it is clean and safe to use for training models
- Why is this needed?
- How many times we really go through the data in detail and rather depends on a data story board!
- Using the data for future growth
- Acting cautiously



Data Roles

Key Roles

Data Analyst

- Collect, clean data from various sources
- Exploratory analysis
- Report Preparation

Data Engineer

- Manage DBMS
- Develop & maintain data pipelines

Data Scientist

- Develop & maintain statistical & ML Models
- Conduct analysis for deep insights

Data Architect

- Create & maintain organisations data blueprint and data systems

DB Admin

- Install & manage DB systems on infra

BI Analysts

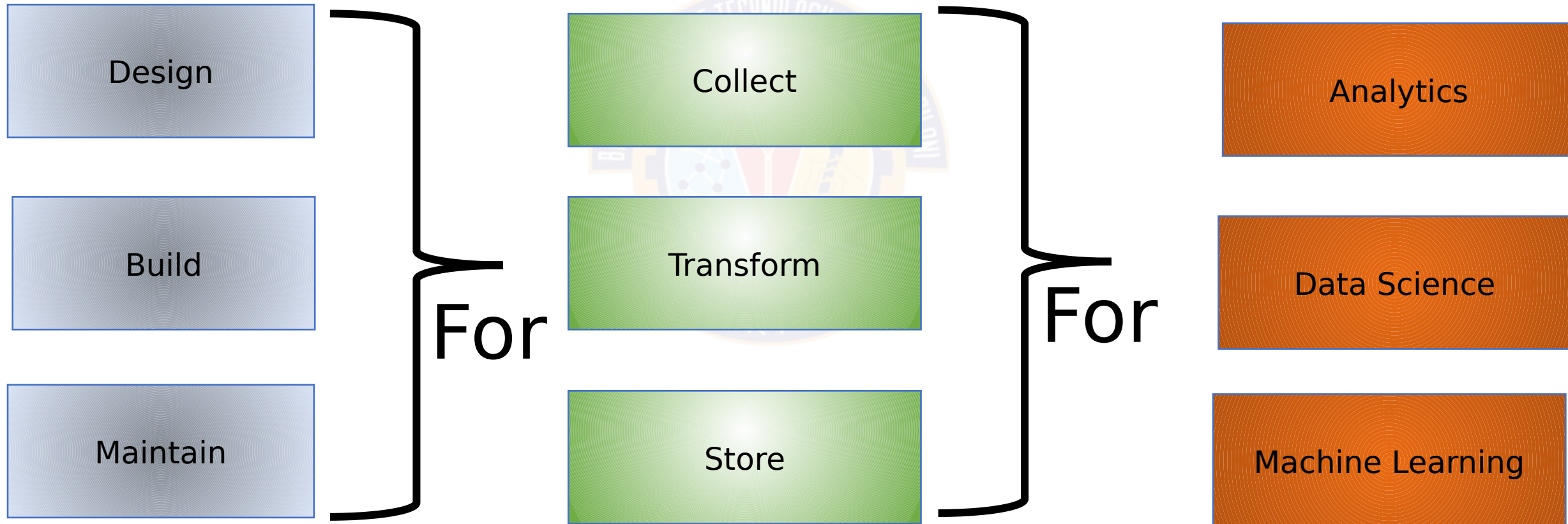
- Focus on business insights and reporting



Connecting Dots

Data Engineering

- From our primary understanding of data management - this is the first concept that we need to understand
- Let us reiterate what we do here?



Case Study

Consider an ecommerce company which wants to understand the customers' shopping interest and want to implement personalized recommendation engine

- Can you say how will you start and what will you do in this scenario?

Data Engineering

- Look into data sources, cleaning logic, storage
- Schedulers to collect the data and feed the data to the recommendation engine
- Maintain
- Raise alarms
- Work on CRs if there are any changes

What is/are missing here?

What is/are missing here?

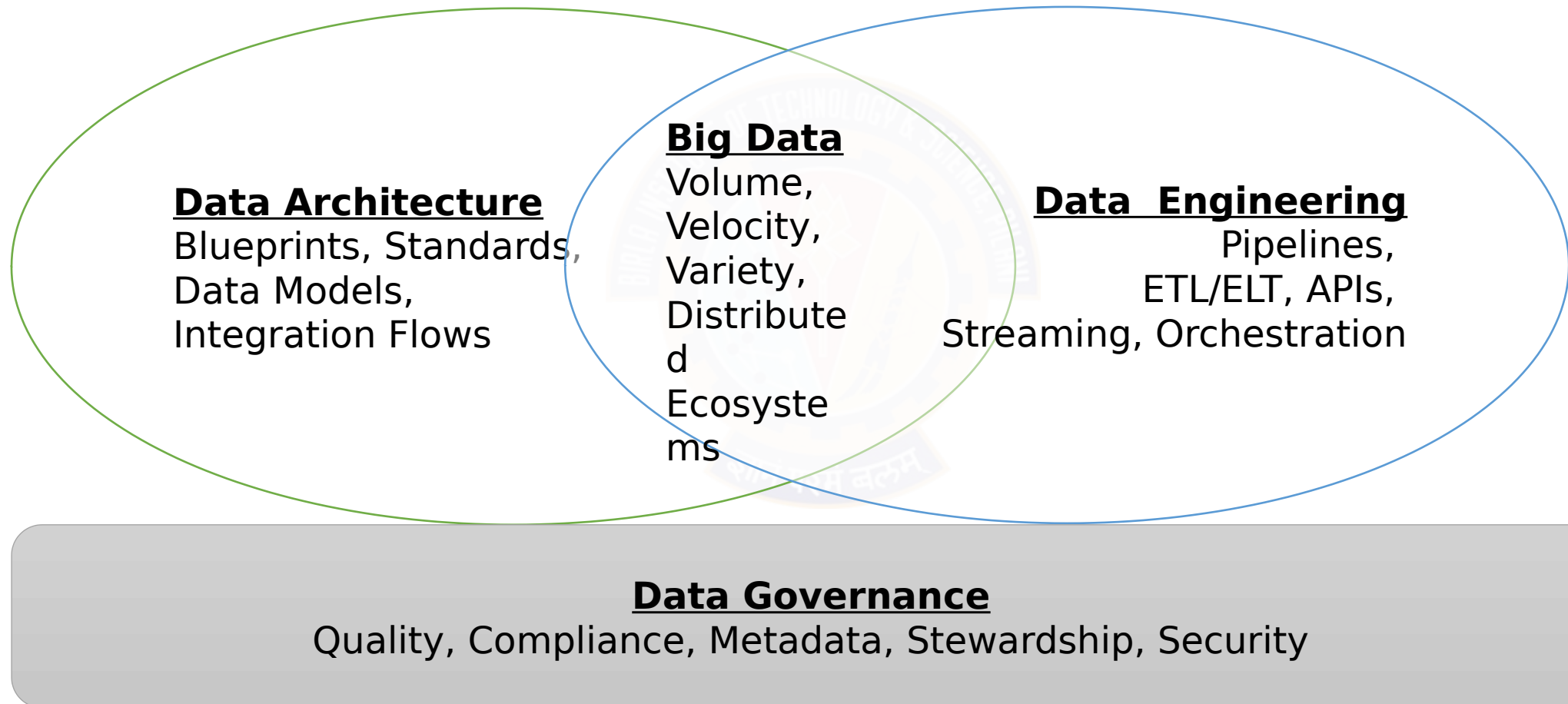
- Business acumen
- Data models
- Holistic view
- Design
- Orchestration
- Overall data platform

Data Architecture

- Privacy
- Security
- Monitoring & logging
- Lifecycle

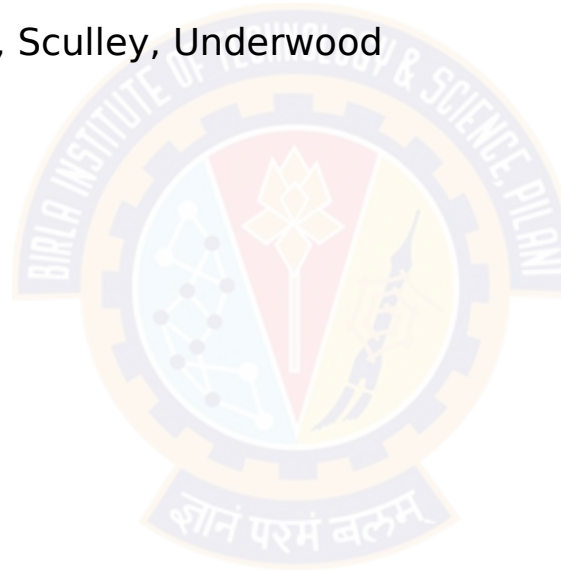
Governance

Conceptual Relationship



References:

- [Source : trellance](#)
- <https://www.ibm.com/in-en/topics/data-management>
- https://medium.com/@icaro_vazquez/what-is-a-data-platform-081549a2cef6
- DMBOK
- Reliable Machine Learning By Chen, Murphy, Sculley, Underwood





Thank You!